Title
MetaCyto: A Tool for Automated Meta-analysis of Mass and Flow Cytometry Data

Journal
Cell Reports, 24(5)

ISSN
2639-1856

Authors
Hu, Zicheng
Jujjavarapu, Chethan
Hughey, Jacob J
et al.

Publication Date
2018-07-01

DOI
10.1016/j.celrep.2018.07.003

Peer reviewed

HHS Public Access

Author manuscript

Cell Rep. Author manuscript; available in PMC 2019 June 19.

Published in final edited form as:

Cell Rep. 2018 July 31; 24(5): 1377–1388. doi:10.1016/j.celrep.2018.07.003.

MetaCyto: A tool for automated meta-analysis of mass and flow cytometry data

Zicheng Hu[1], **Chethan Jujjavarapu**[1], **Jacob J. Hughey**[2], **Sandra Andorf**[3], **Hao-Chih Lee**[4], **Pier Federico Gherardini**[5], **Matthew H. Spitzer**[5,6,7,8], **Cristel G Thomas**[9], **John Campbell**[9], **Patrick Dunn**[9], **Jeff Wiser**[9], **Brian A. Kidd**[4], **Joel T. Dudley**[4], **Garry P. Nolan**[10], **Sanchita Bhattacharya**[1,11], and **Atul J. Butte**[1,11,12,*]

[1]Institute for Computational Health Sciences, University of California, San Francisco, San Francisco, CA, 94158,USA

[2]Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN, 37203, USA

[3]Sean N. Parker Center for Allergy and Asthma Research at Stanford University, Stanford, CA 94305, USA

[4]Institute for Next Generation Healthcare, Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, 10029,USA

[5]Parker Institute for Cancer Immunotherapy, San Francisco, CA, 94129, USA

[6]Department of Microbiology and Immunology, University of California, San Francisco, CA, 94143, USA

[7]Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, CA, 94143, USA

[8]Department of Otolaryngology, University of California, San Francisco, CA, 94143, USA

[9]Northrop Grumman Technology Services Health IT, Rockville, MD, 20850, USA

[10]Department of Microbiology and Immunology, Stanford University, Stanford, CA, 94305, USA

[11]These authors contributed equally

[12]Lead Contact

SUMMARY

[*]Correspondence should be addressed to A.J.B. (Atul.Butte@ucsf.edu).

AUTHOR CONTRIBUTIONS

Z.H., A.J.B. and S.B. conceived the study and developed the method. Z.H. and C.J. designed the overall structure of MetaCyto and performed the meta-analysis of cytometry data from ImmPort. J.H., M.S., P.F.G., S.A., P.D., C.T., J.W., G.N. gave valuable input and suggestions for analysis. H.L., B.A.K. and J.T.D. evaluated the performance of ACDC. Z.H. and C.J. wrote the manuscript and made figures with input from co-authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Code availability

The MetaCyto R package is available on Bioconductor: bioconductor.org/packages/MetaCyto

The source codes of the analysis are available on GitHub: github.com/hzc363/MetaCyto_Paper_Code

While meta-analysis has demonstrated increased statistical power and more robust estimations in studies, the application of this commonly accepted methodology to cytometry data has been challenging. Different cytometry studies often involve diverse sets of markers. Moreover, the detected values of the same marker are inconsistent between studies due to different experimental designs and cytometer configurations. As a result, the cell subsets identified by existing auto-gating methods cannot be directly compared across studies. We developed MetaCyto for automated meta-analysis of both flow and mass cytometry (CyTOF) data. By combining clustering methods with a silhouette scanning method, MetaCyto is able to identify commonly labeled cell subsets across studies, thus enabling meta-analysis. Applying MetaCyto across a set of ten heterogeneous cytometry studies totaling 2,926 samples enabled us to identify multiple cell populations exhibiting differences in abundance between demographic groups. Software is released to the public through Bioconductor (bioconductor.org/packages/MetaCyto).

INTRODUCTION

Meta-analysis of existing data across different studies offers multiple benefits. The aggregated data allows researchers to test hypotheses with increased statistical power. The involvement of multiple independent studies increases the robustness of conclusions drawn. In addition, the complexity of aggregated data allows researchers to test or generate new hypotheses. These benefits have been shown by many studies in areas such as genomics, cancer biology and clinical research and have led to important new biomedical findings (Boulé et al., 2001; Kodama et al., 2012; Sutton et al.; Wirapati et al., 2008). For example, one study showed the correlation between neo-antigen abundance in tumors and patient survival by performing meta-analysis of RNA sequencing data from The Cancer Genome Atlas (TCGA) (Brown et al., 2014). In another study, meta-analysis of genome-wide association studies identified novel loci that affect the risk of type 1 diabetes (Barrett et al., 2009).

With the recent advances in high-throughput cytometry technologies the immune system can be characterized at the single cell level with up to 45 parameters, minimizing the technical limitations and allowing capture of more valuable information from immunology studies (Bandura et al., 2009; Perfetto et al., 2004; Shapiro, 1983). Open science initiatives have led to more of this type of research data being accessible, and the availability of shared cytometry data, including data from flow cytometry and mass cytometry (CyTOF), is growing exponentially. Notably, the ImmPort database (www.immport.org), a repository for immunology-related research and clinical trials, provides numerous studies with thousands of cytometry datasets (Bhattacharya et al.). However, meta-analysis of cytometry datasets remains particularly challenging. Different studies use diverse sets of protein markers and fluorophore/isotope combinations. The detected values of the same marker are inconsistent between studies because of different cytometer configurations or operators. In addition, the high dimensionality of cytometry data, especially CyTOF data, makes manual gating based meta-analysis difficult and time-consuming.

Multiple computational methods have been proposed to automate the analysis of cytometry data, such as FlowSOM (Van Gassen et al., 2015), FlowMeans (Aghaeepour et al., 2011)

and CITRUS (Bruggner et al., 2014). Although they are effective in analyzing data from a single study (Aghaeepour et al., 2013; Weber and Robinson, 2016), several limitations have prevented their use in meta-analysis. First, the results of these methods cannot be directly compared across studies. The cell subsets identified by these methods are usually labeled with anonymous identifiers with no cell-specific annotation, making it impossible to identify common cell populations across different studies. In addition, many clustering methods are sensitive to parameter choices. For example, FlowSOM, FlowMeans and SPADE (Qiu et al., 2011) require users to pre-specify the number of clusters. As a result, extensive parameter tuning and manual inspection are required for every cytometry dataset. In meta-analysis where large numbers of input datasets could be involved, these manual selected choices become a major technical burden.

In this study, we developed MetaCyto to enable automated meta-analysis of cytometry datasets, including data from both conventional flow and CyTOF cytometry data. MetaCyto is able to accurately identify common cell populations across studies without parameter tuning requirements. It then applies hierarchical models to robustly estimate the effects of factors of interest, such as age, ethnicity or vaccination, on the cell populations using data across all input studies.

To test the utility of MetaCyto, we performed a joint analysis of 10 human immunology cytometry datasets contributed by four different institutions (Ledgerwood et al., 2012; Obermoser et al., 2013; Wertheimer et al., 2014; Whiting et al., 2015). Altogether, this analysis spanned 2926 whole blood or peripheral blood mononuclear cells (PBMC) samples from 984 healthy subjects, which were acquired using either flow cytometry or CyTOF with a diverse set of markers. Among these 984 subjects, over 90 percent were White or Asian. While it is well known that characteristics of multiple immune system-related diseases, such as HIV(Achhra et al., 2010), systemic lupus erythematosus (Petri, 2002) and hepatitis C (Golden-Mason et al., 2008), vary between the two ethnic groups, the heterogeneity of the immune system among the human population has made studying these differences difficult (Brodin et al., 2015; Li et al., 2016). We hypothesized that a meta-analysis approach could lead to a better understanding of differences in the immune system between ethnic groups. Using MetaCyto, we not only confirmed a known difference, but also identified new cell types whose frequencies differ between White and Asian.

RESULTS

MetaCyto identifies common cell subsets across studies

Our meta-analysis of cytometry data follows four steps: data aggregation, data pre-processing, identification of common cell subsets across studies, and statistical analysis (Figure 1A). The third step, identification of common cell subsets across studies, has been one of the main technical challenges preventing automated meta-analysis. Therefore, while all four steps are automated and covered in the MetaCyto software system and documented in the online methods, here we primarily focus on describing our identification and relating of common cell subsets across studies.

MetaCyto employs two automated pipelines, unsupervised analysis and guided analysis, to identify common cell subsets across studies. The unsupervised analysis pipeline identifies cell subsets in a fully automated way. Cytometry data in each study is first clustered using an existing clustering method (Figure 1B **Top**). FlowSOM (Van Gassen et al., 2015) was implemented as the default clustering method due to its speed and performance. However, any other clustering method, such as hierarchical clustering or FlowMeans, could be substituted as well. At this stage, clusters are labeled with non-informative labels, such as C1, C2, C3, which cannot be related across studies. For example, C1 in study 1 and C1 in study 2 represent entirely different cell populations.

A threshold is then chosen to bisect the distribution of each marker into positive and negative regions, needed to label each cluster in a biologically meaningful way (Figure 1B **Middle**). The selection of a threshold is easy when a clear bi-modal distribution is present, but becomes challenging in other cases. We implemented a Silhouette scanning method, which bisects each marker at the threshold maximizing the average silhouette, a widely used way of describing the quality of clusters (Rousseeuw, 1987). We compared Silhouette scanning against 8 other bisection methods and found it to be superior (Figure S1 and Table S1).

Clusters are then labeled for each of the markers based on the following two rules: first, if the marker levels of 95% of cells in the cluster are above or below the threshold, the cluster will be labeled as positive or negative for the marker, respectively. Otherwise, the cluster will not be labeled for the marker. For example in Figure 1B, both C2 and C1 in study 2 will be labeled as CD8+ CD4-; second, if a marker is positive or negative in 95% of all cells, the marker is not used to label any clusters. For example, CD45, which is expressed by all immune cells, will not be used to label any cell clusters in the blood. The two rules are used to reduce redundancy and ensure that only the informative markers are used for labeling.

Next, clusters with the same labels are merged into a square shaped cluster (Figure 1B **Bottom**). In cytometry data with higher dimensions, clusters are hyper-rectangles. Following this stage, common cell subsets across studies can be rigorously identified and annotated. For example, the CD4- CD8+ clusters in both study 1 and study 2 correspond to CD8+ T cells.

The guided analysis pipeline identifies cell subsets using pre-defined cell definitions, thus allowing for the search of specific cell subsets defined by immunologists. After bisecting each marker into positive and negative regions, cells fulfilling the pre-defined cell definitions are identified. For example, the CD3+ CD4+ CD8- (CD4+ T cells) cell subset corresponds to the cells that fall into the CD3+ region, CD4+ region and CD8- region concurrently (Figure 1C). Notice that both CD45RA+ and CD45RA- populations are included in the cell subset, because the cell definition does not specify the requirement for CD45RA expression. However, researchers could easily alter the cell definition to CD3+ CD4+ CD8- CD45RA+ to find the CD45RA+ cell subset.

Evaluating the guided analysis pipeline

A successful meta-analysis of cytometry data requires cell populations to be identified accurately from each study. To evaluate if the guided analysis pipeline of MetaCyto can accurately identify cell subsets from a single study, we downloaded a set of PBMC cytometry data (SDY478) from ImmPort, with which the original authors identified 88 cell types. Correspondingly, we specified the 88 cell definitions (Table S2) based on the author's gating strategy and identified these cell subsets for each cytometry sample using the guided analysis pipeline in MetaCyto. We compared the proportions of all cell subsets estimated by MetaCyto with the original manual gating results and found that MetaCyto estimations are highly consistent with the manual gating result (Figure 2A-C). We compared our estimations to two existing methods, flowDensity (Malek et al., 2015) and ACDC (Lee et al., 2017), which can also identify pre-defined cell populations. Our results suggest that MetaCyto's quantification of both major and rare populations were more accurate than FlowDensity's (Figure 2D,E). Although ACDC and MetaCyto results had the same correlation with manual gating, ACDC tended to over-estimate the cell abundance (Figure S2A,B). In addition, a relatively shorter computational time of MetaCyto (around 3 minutes) compare to ACDC (over 2 hours) makes it advantageous in analyzing a large number of datasets.

Evaluating the unsupervised analysis pipeline

We then tested the performance of the unsupervised analysis pipeline of MetaCyto. In the unsupervised analysis pipeline, cell clusters are first identified by an existing clustering algorithm. The subsets are then labeled using informative markers and are merged into hyper-rectangle clusters based on the labeling result (Figure 1B). To learn how such a merge affects the quality of clusters, we evaluated the results of two clustering algorithms, FlowSOM (Van Gassen et al., 2015) and FlowMeans (Aghaeepour et al., 2011), with and without the merging step. Multiple studies have been conducted to evaluate the performance of existing clustering method for cytometry data (Aghaeepour et al., 2013; Weber and Robinson, 2016). The most recent (Weber and Robinson, 2016) compared 15 clustering methods and found FlowSOM generally outperformed other methods after manual tuning.

We downloaded an evaluation dataset, West Nile virus dataset (FlowCAP WNV), used by Weber et al (Weber and Robinson, 2016) and applied FlowSOM. The clustering result is then labeled and merged. Since FlowSOM requires a pre-specified cluster number (K), we did multiple runs with K ranging from 10 to 90. F-measure is used to evaluate the quality of the clusters. We found that the quality of clusters is comparable before and after merging when K equals to 10. However, the performance of FlowSOM drops when K increases. The subsequent merging step prevented FlowSOM performance to deteriorate (Figure 2F). We then looked at the total number of clusters identified before and after merging. As expected, FlowSOM identified the same number of clusters specified by K. However, when running the merging step after FlowSOM, the total number of clusters no longer increases after a certain point (Figure 2G). The same results were obtained with FlowMeans (Aghaeepour et al., 2011) (Figure S2C,D).

To see if such benefit of the merging step only exists in datasets where the intrinsic number of cell subsets is small, we applied the same methodology in the normal donor (ND) dataset

from FlowCAP competition (Aghaeepour et al., 2013), where more cell subsets can be identified. Consistent with results from WNV dataset, the merging step is able to prevent the over-partitioning in the ND dataset as well (Figure S2E-F).

The results suggest that MetaCyto is able to merge small clusters in a biologically meaningful way, preventing over-partitioning of the cell subsets, thus allowing the clustering analysis to be performed without tuning any parameters.

Meta-analysis using MetaCyto confirms previous findings

After evaluating the performance of MetaCyto in analyzing cytometry data from single studies, we next tested the ability of MetaCyto in yielding consistent results from combining multiple studies. We applied MetaCyto to identify cell types whose frequencies are different between age, gender and ethnic groups. We downloaded 10 studies from ImmPort containing cytometry data. These ten studies had been contributed from four different institutions, where 86 panels containing 74 different markers were used (Figure 3 **and** Table S3). Altogether, the datasets contain 2926 whole blood or PBMC samples from 984 healthy subjects and were acquired using either flow cytometry or CyTOF. We obtained the demographic information, including age, gender and ethnicity, directly from the metadata associated with the studies. The subjects are proportionately distributed by gender, with slightly more female than male. The age span ranges from 19 to 90 years. The subjects come from five different defined ethnic groups. Among them, over 90% were White or Asian (Figure S3).

We used both unsupervised and guided MetaCyto analysis pipelines in parallel to identify cell subsets. For the latter, we used 23 cell type definitions from the Human ImmunoPhenotyping Consortium (HIPC) (Finak et al., 2016), ranging from effector memory T cells to monocytes (Table S4). We then estimated the effect size of age, gender and ethnicity on the cell type proportions using hierarchical statistical models.

Because the 10 studies differ in multiple aspects, including the sample size, the cytometry experimental design (Figure 3) and the distribution of demographics (Figure S3), it is important to determine if results from these studies can be combined in a meta-analysis. We first performed a ten-fold leave-one-out analysis. Each time, we left one of the 10 studies out and estimated the effect sizes of age, gender and ethnicity using the rest nine studies. We found that the leave-one-out analysis agree well with the full meta-analysis (correlation ranges from 0.76 to 1, Table S5), suggesting that the meta-analysis results are not dominated by one study. In addition, we performed Cochran's Q tests on the results from 10 studies. The tests did not identify significant heterogeneity between studies (p values range from 0.22 to 1).

We then validated our results using the effect sizes of age and gender, previously well characterized in other studies (Carr et al., 2016; Whiting et al., 2015). We tested whether results obtained with MetaCyto could replicate results from a previous independent study (Carr et al., 2016). Among the 23 cell types identified by the guided analysis pipeline, 14 overlapped with the cell types included in the Carr study. We compared the effect size of age and gender on the proportion of these 14 cell types, between MetaCyto on the 10 studies,

and the independent results from Carr, et al. We found that results agree well with each other on both the effect size of age (r = 0.69, p = 0.006, Figure 4A) and gender (r = 0.71, p = 0.004, Figure 4B). The result, together with results from the leave-one-out analysis and Cochran's Q tests, suggest that data from the 10 studies can be analyzed together in a meta-analysis using MetaCyto.

The only discrepancy between our analysis and Carr study was the effect of age on CD8+ T cells (Figure 4A). Our result showed that the proportion of CD8+ T cells significantly decreases with age, while Carr study reported an increase with age. We visually inspected MetaCyto's auto-gating, and ruled out such disagreement being caused by gating errors in our study (Figure 4C). The forest plot showed that our finding was consistent across cytometry panels (Figure 4D). In the literature, one study found that CD8+ T cell proportion decrease with age (Yan et al.) while another study found no association between CD8+ T cells and age (Uppal et al., 2003). These discrepancies suggest that the effect of age on CD8+ T cells is highly variable and environment specific factors might be contributing to these results. Future studies are needed to identify the exact factors.

Meta-analysis using MetaCyto identifies previously unreported differences in immune cells between ethnic groups

Our meta-analysis using the guided pipeline in MetaCyto revealed five cell types to be significantly different between the Asian and White. Asians have higher percentages of total CD4+ T cells and CD4+ central memory T cells, and lower percentages of natural killer (NK) cells, naïve CD8+ T cells and total CD8+ T cells (Figure 5A). Among these findings, only the difference of total CD4+ T cells has been reported previously(Howard et al., 1996). MetaCyto was able to identify this ethnic difference consistently across all cytometry panels (Figure 5B). Combining the results from all panels allowed us to confirm the difference with high confidence ($p = 1.2\times10^{-7}$).

In all ten studies, Asian individuals make up less than 25% of the cohorts. To test if our findings are affected by the data imbalance, we down-sampled White individuals so that the number of White and Asian individuals are equal. We found that the effect sizes are consistent before and after down sampling (correlation = 0.92). Importantly, the same ethnic differences (CD4+ T cells, NK cells, naïve B cells, CD4 central memory cells and CD8 T cells) are observed after down sampling (Figure S4).

To further confirm the four previously unreported ethnic differences, we inspected the results from MetaCyto in detail. First, we visualized the identified cell populations in all studies, and confirmed that our results were not artifacts of automated gating (Figure 6 A-D). Second, as described in the previous section, we tested if these ethnic differences were consistent across cytometry panels. Cochran's Q test did not identify significant heterogeneity between cytometry panels (p-values equal to 0.86, 0.27, 0.71 and 0.90 for NK cells, naïve B cells, CD4 central memory cells and CD8 T cells respectively). Visual inspection of the forest plots also confirmed that the results were consistent in most of the cytometry panels (Figure 6E-H).

Results from the unsupervised analysis identified multiple cell types, other than the 23 types used in the guided analysis, whose abundance were different between Asian and White (Table S6). As one example, we found that the proportion of a sub-population of CD8+ T cells, the CD3+ CD4- CD45RA+ CD8+ CD85J- cell population, is significantly higher in Asians than in Whites (Figure S5). A closer look at the forest plot revealed that the association between this population and ethnicity was not at a significant level in most studies taken independently. However, by combining the results from multiple studies, we were able to identify this association with high confidence (p=0.0049).

DISCUSSION

With the collection of publically available cytometry studies rapidly growing, researchers can often identify multiple studies that were designed or can be re-purposed to answer a common research question. Meta-analysis of these studies allows researchers to answer the research question with a more robust conclusion and higher statistical power. Many cytometry studies that are publically available include hundreds of high dimensional cytometry data. Performing meta-analysis manually on these studies is not only time consuming, but also prone to human error and bias. In this study, we developed and demonstrated a computational tool called MetaCyto, which allows fully automated meta-analysis of both CyTOF and flow cytometry data.

When performing a meta-analysis of cytometry data, a big challenge lies in the identification of common cell subsets across heterogeneous cytometry studies. In MetaCyto, we implemented two complementary analysis pipelines to automate the cell identification process. The guided analysis pipeline is able to identify cell populations using user-defined cell definitions. For example, regulatory T cells can easily be identified using the definition “CD3+ CD4+ Foxp3+”. Such an approach allows researchers to incorporate their domain knowledge into the analysis, making the result more biologically relevant. In addition to the guided analysis pipeline, MetaCyto also allows researchers to identify cell populations in an unsupervised manor. Due to the high dimensionality of cytometry data, an exhaustive grid search will lead to an astronomical number of cell subsets. For example, if we divide each marker into positive and negative regions, 45 markers in a CyTOF experiment have 2^{45} combinations. To avoid such a situation, the unsupervised pipeline in MetaCyto first identifies cell clusters using a clustering method. Successful efforts were made by the community to develop efficient clustering methods for flow cytometry data analysis. We built MetaCyto to be fully compatible with existing clustering methods. MetaCyto is able to merge and transform the clusters from existing clustering algorithms in a biologically meaningful way, therefore improving result quality and enabling further meta-analysis of many studies.

Based on the test result, we recommend over-clustering the data first, followed by the merging of the clusters by MetaCyto. Such a strategy not only makes the method tuning free, but also is more computationally efficient than traditional auto-tuning methods, which require running the clustering algorithm multiple times with different parameters.

In MetaCyto, the distribution of each marker is bisected into positive and negative regions using a silhouette-scanning method. However, some markers may show tri-modal distributions. Although the silhouette-scanning method can easily be modified to divide the distribution into three regions (low-medium-high), only bisection is used in MetaCyto for the following reasons. First, it is known that multiple technical factors, such as auto-fluorescence, compensation, transformation and non-specific binding of antibodies, can lead to false tri-modal distributions (Morice et al., 2004; Ray and Pyne, 2012). In these cases, the low-medium-high regions do not represent distinct cell populations. Second, upon examining multiple cytometry studies, we found that although some markers (e.g. CD8, CD45RA, CD127) show tri-modal distributions in certain cytometry studies, they show bi-model distributions in other studies. Such inconsistency makes it difficult to reliably relate cell subsets across studies. Finally, our test result shows that bisection using silhouette scanning is able to identify the population that is truly positive for a marker even when the distribution is not bi-modal.

It should be noted that abnormally “bright” particles, such as beads and dead cells, will affect the silhouette scanning method. Therefore, we recommend gating out the “bright” particles before performing the meta-analysis. The MetaCyto R package allow users to perform pre-gating using a user defined strategy, such as “PI- FSC+” for flow cytometry data and “Bead- DNA+” for CyTOF data.

A recent study (Diggins et al., 2017) have proposed a novel method to annotate cell subsets using maker enrichment modeling (MEM) scores. Although the approach is highly effective in individual datasets, several limitations exist for its use in meta-analysis. First, the enrichment score is context dependent and varies between studies, making it difficult to identify common cell types across studies. Second, the MEM method is designed to label cell populations rather than identifying cell populations. As a result, if a clustering algorithm identifies a cell subset in dataset 1 but not in dataset 2, the cell subset will be missed in a meta-analysis. In contrast, MetaCyto identifies the cell subset in both datasets, allowing meta-analysis.

By combining samples from multiple studies, meta-analysis is able to increase the statistical power of hypothesis testing. One concern is that such approach may reach significant p values of very weak biological phenomenon. Therefore, We would always encourage users of MetaCyto to not just look at the statistics significance, but also the effect size. Our meta-analysis identified 4 ethnic differences in immune cells, which to our knowledge have not been reported previously. The findings not only have significant p values, but also have large effect sizes (around 0.3). The effect sizes are comparable with the effect size of CD4 T cells, a well-characterized ethnic difference (Howard et al., 1996), suggesting that these findings reflect important biological differences in the immune system.

There are several potential limitations of the current study. In the unsupervised analysis pipeline of MetaCyto, although the merging step makes the clustering result more robust, it may eliminate some small cell populations of biological meaning. To overcome this limitation, researchers can use a more sensitive method, such as CITRUS (Bruggner et al., 2014), to identify the cell subsets of interest from a single study. They can then craft cell

definitions for those subsets and use the guided analysis pipeline of MetaCyto to perform meta-analysis across studies. Another limitation is that our meta-analysis only established correlations, rather than causations, between cell populations and ethnicity. In-depth studies are needed to further validate our findings and to identify the genetic or environmental causes of these differences.

In summary, we developed MetaCyto, a computational tool that allows the automated meta-analysis of cytometry data. Applying MetaCyto to cytometry data from 10 human immunology studies allowed us to thoroughly characterize differences in the immune system between Asian and White populations. Other than the previously known differences in CD4+ T cell abundance, we identified previously unreported cell populations whose abundance were significantly different between the two ethnicities, and demonstrated that the findings are consistent across multiple independent studies. Our findings can help us better understand the heterogeneity of the human immune system in the population. They also serve as a starting point for future studies to reveal the mechanisms behind ethnic discrepancies in immune-related diseases

EXPERIMENTAL PROCEDURES

Data Aggregation

Flow cytometry data and CyTOF data from SDY112, SDY167 (Ledgerwood et al., 2012), SDY180 (Obermoser et al., 2013), SDY311, SDY312, SDY314, SDY315, SDY420 (Whiting et al., 2015), SDY478 and SDY736 (Wertheimer et al., 2014) were downloaded from ImmPort web portal. Only fcs files from pre-vaccination blood samples of healthy adults were included in the meta-analysis. Parameters, including antibodies and fluorescence or isotope labels, used in each fcs file were then identified using the *fcsInfoParser* function in MetaCyto. The fcs files were then organized into panels, which are defined as a collection of fcs files from the same study that have the same set of parameters.

We obtained the demographic information directly from the metadata associated with each study. Specifically, the age, gender and ethnicity information were obtained from the “Subject_2_Flow_cytometry_result.txt” or “Subject_2_CyTOF_result.txt” tables. The ethnicity categories were standardized according to the Standards for the Classification of Federal Data on Race and Ethnicity (Federal Registrar, 1997) by the ImmPort data curation team.

Manual gating results for both FlowCAP WNV data (ID number FR-FCM-ZZY3) were downloaded from the FlowRepository link: community.cytobank.org/cytobank/experiments/4329.

All data sets were downloaded between September 1, 2016 and February 1, 2017.

Data Pre-processing

Flow cytometry data from ImmPort were compensated for fluorescence spillovers using the compensation matrix supplied in each fcs file. All data from ImmPort were arcsinh transformed. For flow cytometry data, the formula f(x) = arcsinh (x/150) was used. For

CyTOF data, the formula f(x) = arcsinh (x/8) was used. All transformation and compensation were done using the *preprocessing* or *preprocessing.batch* function in MetaCyto.

Cytometry data FlowCAP WNV was transformed and subset to only include protein markers. The pre-processing was doing using the same code provided by the Weber study (Weber and Robinson, 2016) : github.com/lmweber/cytometry-clustering-comparison

Bisecting Marker Distributions using Silhouette scanning

The range of a marker was divided into 100 intervals using 99 breaks. The distribution was bisected at each break and the corresponding average silhouette (Rousseeuw, 1987) was calculated. The break giving rise to the largest average silhouette was used as the cutoff for bisection.

Identifying cell subsets with the guided analysis pipeline in MetaCyto

Cell definitions were created based on the gating strategies provided by authors of SDY 420 and SDY478 from ImmPort database or based on the cell definition from the Human ImmunoPhenotyping Consortium (Finak et al., 2016). The cell definitions are available in the Table S1, S2, S4.

To identify the corresponding cell subsets, Silhouette scanning was used to bisect the distribution of cell markers into positive and negative regions. Cells fulfilling the cell definitions were then identified. For example, the CD3+ CD4+ CD8- (CD4+ T cells) cell subset corresponds to the cells that fall into the CD3+ region, CD4+ region and CD8- region concurrently. The proportion of each cell subset in blood was calculated by dividing the number of cells in the subset by the total number of cells in the blood. The procedure is performed using the *searchCluster* or *searchCluster.batch* function in the MetaCyto package.

Identifying cell subsets with the unsupervised analysis pipeline in MetaCyto

FlowSOM (Van Gassen et al., 2015) or FlowMeans (Aghaeepour et al., 2011) were used to identify cell clusters in the cytometry data. Silhouette scanning was used to identify a threshold that bisects the distribution of cell markers into positive and negative regions. To label the identified cell clusters, the marker levels in each cluster were compared with the bisection threshold. If the marker levels of 95% of cells in the cluster are above or below the threshold, the cluster will be labeled as positive or negative for the marker, respectively. Otherwise, the cluster will not be labeled for the marker. If a marker is positive or negative in 95% of all cells, the marker is not used to label any clusters. The procedure is performed using the *labelCluster* function in MetaCyto.

MetaCyto then identifies the corresponding cell subsets using the generated labels, in a fashion similar to the guided analysis pipeline. Notice that such a process is equivalent of merging cell clusters that have the same labels into a hyper-rectangle shaped cluster. To capture all the identified cell subsets, the MetaCyto pools the labels from different studies and quantifies the corresponding cell subsets in all studies, as long as the studies contain the necessary cell marker. The proportion of each cell subset in blood was calculated by dividing

the number of cells in the subset by the total number of cells in the blood. The procedure was performed using the *searchCluster* or *searchCluster.batch* function in the MetaCyto package.

Statistical Analysis

2-level hierarchical regression models were used in the meta-analysis of the 10 human immunology studies from ImmPort: the proportion of cell subsets was regressed against age, gender and ethnicity (Y ~ age + gender + ethnicity) in each cytometry panel. The effect size was defined as the regression coefficient divided by the standard deviation of Y. The overall effect size from all cytometry panels was estimated using a random effect model. For data from the Carr study, the proportion of a cell population was regressed against age and gender. Ethnicity information was missing in the data, therefore was omitted in the regression. All statistical analysis was performed using the *metaAnalysis* function in MetaCtyo. The p-value was adjusted using the Benjamini-Hochberg (Author et al., 1995) correction.

To test the heterogeneity in Meta-analysis, Cochran's Q test was performed using the *cochran.Q* function in the *Mada* package.

In Figure 4A and B, Pearson correlations are calculated and tested against the null hypothesis (correlation equals zero) using the *cor.test* function in R.

In Figure S5B, Shapiro-Wilk test was performed to check the normality assumption using the *shapiro.test* function in R. F test was performed to check the equal variance assumption using the *var.test* function in R. A two-sided unpaired Mann-Whitney test is performed to test the difference between two groups using the *wilcox.test* function in R.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank Mark Davis, Julie Ledgerwood, Karolina Palucka, Charles Fathman and Janko Nikolich-Zugich for sharing the data on ImmPort. We thank Marina Sirota, Dvir Aran, Henry Schaefer, Elizabeth Thomson, Kelly Zalocusky and Matthew Kan for helpful discussion.

Research reported in this publication was supported by the National Institute of Allergy and Infectious Diseases (Bioinformatics Support Contract HHSN272201200028C). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES:

Achhra AC, Zhou J, Dabis F, Pujari S, Thiebaut R, Law MG, and Bonnet F (2010). Difference in absolute CD4+ count according to CD4 percentage between Asian and Caucasian HIV-infected patients. J. AIDS Clin. Res. 1, 1–4. [PubMed: 21479149]

Aghaeepour N, Nikolic R, Hoos HH, and Brinkman RR (2011). Rapid cell population identification in flow cytometry data. Cytom. Part A 79A, 6–13.

Aghaeepour N, Finak G, Dougall D, Khodabakhshi AH, Mah P, Obermoser G, Spidlen J, Taylor I, Wuensch SA, Bramson J, et al. (2013). Critical assessment of automated flow cytometry data analysis techniques. Nat. Methods 10, 228–238. [PubMed: 23396282]

Author T, Benjamini Y, Hochberg Y, and Benjaminit Y (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. Source J. R. Stat. Soc. Ser. B J. R. Stat. Soc. Ser. BMethodological) J. R. Stat. Soc. B 57, 289–300.

Bandura DR, Baranov VI, Ornatsky OI, Antonov A, Kinach R, Lou X, Pavlov S, Vorobiev S, Dick JE, and Tanner SD (2009). Mass Cytometry: Technique for Real Time Single Cell Multitarget Immunoassay Based on Inductively Coupled Plasma Time-of-Flight Mass Spectrometry. Anal. Chem. 81, 6813–6822. [PubMed: 19601617]

Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, Erlich HA, Julier C, Morahan G, Nerup J, Nierras C, et al. (2009). Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat. Genet. 41, 703–707. [PubMed: 19430480]

Bhattacharya S, Andorf S, Gomes L, Dunn P, Schaefer H, Pontius J, Berger P, Desborough V, Smith T, Campbell J, et al. ImmPort: disseminating data to the public for the future of immunology. Immunol. Res. 58, 234–239. [PubMed: 24791905]

Boulé NG, Haddad E, Kenny GP, Wells GA, and Sigal RJ (2001). Effects of Exercise on Glycemic Control and Body Mass in Type 2 Diabetes Mellitus. JAMA 286, 1218. [PubMed: 11559268]

Brodin P, Jojic V, Gao T, Bhattacharya S, Angel CJL, Furman D, Shen-Orr S, Dekker CL, Swan GE, Butte AJ, et al. (2015). Variation in the Human Immune System Is Largely Driven by Non-Heritable Influences. Cell 160, 37–47. [PubMed: 25594173]

Brown SD, Warren RL, Gibb EA, Martin SD, Spinelli JJ, Nelson BH, and Holt RA (2014). Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. Genome Res. 24, 743–750. [PubMed: 24782321]

Bruggner RV, Bodenmiller B, Dill DL, Tibshirani RJ, and Nolan GP (2014). Automated identification of stratifying signatures in cellular subpopulations. Proc. Natl. Acad. Sci. 111, E2770–E2777. [PubMed: 24979804]

Carr EJ, Dooley J, Garcia-Perez JE, Lagou V, Lee JC, Wouters C, Meyts I, Goris A, Boeckxstaens G, Linterman MA, et al. (2016). The cellular composition of the human immune system is shaped by age and cohabitation. Nat. Immunol. 17, 461–468. [PubMed: 26878114]

Diggins KE, Greenplate AR, Leelatian N, Wogsland CE, and Irish JM (2017). Characterizing cell subsets using marker enrichment modeling. Nat. Methods 14, 275–278. [PubMed: 28135256]

Federal Registrar (1997). Revisions to the standards for the classification of federal data on race and ethnicity. Regist. Fed.

Finak G, Langweiler M, Jaimes M, Malek M, Taghiyar J, Korin Y, Raddassi K, Devine L, Obermoser G, Pekalski ML, et al. (2016). Standardizing Flow Cytometry Immunophenotyping Analysis from the Human ImmunoPhenotyping Consortium. Sci. Rep. 6, 20686. [PubMed: 26861911]

Van Gassen S, Callebaut B, Van Helden MJ, Lambrecht BN, Demeester P, Dhaene T, and Saeys Y (2015). FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. Cytom. Part A 87, 636–645.

Golden-Mason L, Klarquist J, Wahed AS, and Rosen HR (2008). Cutting edge: programmed death-1 expression is increased on immunocytes in chronic hepatitis C virus and predicts failure of response to antiviral therapy: race-dependent differences. J. Immunol. 180, 3637–3641. [PubMed: 18322167]

Howard RR, Fasano CS, Frey L, and Miller CH (1996). Reference intervals of CD3, CD4, CD8, CD4/CD8, and absolute CD4 values in asian and non-asian populations. Cytometry 26, 231–232. [PubMed: 8889397]

Kodama K, Horikoshi M, Toda K, Yamada S, Hara K, Irie J, Sirota M, Morgan AA, Chen R, Ohtsu H, et al. (2012). Expression-based genome-wide association study links the receptor CD44 in adipose tissue with type 2 diabetes. Proc. Natl. Acad. Sci. U. S. A. 109, 7049–7054. [PubMed: 22499789]

Ledgerwood JE, Hu Z, Gordon IJ, Yamshchikov G, Enama ME, Plummer S, Bailer R, Pearce MB, Tumpey TM, Koup RA, et al. (2012). Influenza virus h5 DNA vaccination is immunogenic by

intramuscular and intradermal routes in humans. Clin. Vaccine Immunol. 19, 1792–1797. [PubMed: 22956656]

Lee H-C, Kosoy R, Becker CE, Dudley JT, and Kidd BA (2017). Automated cell type discovery and classification through knowledge transfer. Bioinformatics 11, 1822–1833.

Li Y, Oosting M, Deepen P, Ricaño-Ponce I, Smeekens S, Jaeger M, Matzaraki V, Swertz M, Xavier R, Franke L, et al. (2016). Inter-individual variability and genetic influences on cytokine responses to bacteria and fungi. Nat. Med. 22, 952–960. [PubMed: 27376574]

Malek M, Taghiyar MJ, Chong L, Finak G, Gottardo R, and Brinkman RR (2015). flowDensity: reproducing manual gating of flow cytometry data by automated density-based cell population identification. Bioinformatics 31, 606–607. [PubMed: 25378466]

Morice WG, Kimlinger T, Katzmann JA, Lust JA, Heimgartner PJ, Halling KC, and Hanson CA (2004). Flow Cytometric Assessment of TCR-V b Expression in the Evaluation of Peripheral Blood Involvement by T-Cell Lymphoproliferative Disorders: A Comparison With Conventional T-Cell Immunophenotyping and Molecular Genetic Techniques. Am. J. Clin. Pathol. 121, 373–383. [PubMed: 15023042]

Obermoser G, Presnell S, Domico K, Xu H, Wang Y, Anguiano E, Thompson-Snipes L, Ranganathan R, Zeitner B, Bjork A, et al. (2013). Systems Scale Interactive Exploration Reveals Quantitative and Qualitative Differences in Response to Influenza and Pneumococcal Vaccines. Immunity 38, 831–844. [PubMed: 23601689]

Perfetto SP, Chattopadhyay PK, and Roederer M (2004). Innovation: Seventeen-colour flow cytometry: unravelling the immune system. Nat. Rev. Immunol. 4, 648–655. [PubMed: 15286731]

Petri M (2002). Epidemiology of systemic lupus erythematosus. Best Pract. Res. Clin. Rheumatol. 16, 847–858. [PubMed: 12473278]

Qiu P, Simonds EF, Bendall SC, Gibbs KD, Bruggner RV, Linderman MD, Sachs K, Nolan GP, and Plevritis SK (2011). Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. Nat. Biotechnol. 29, 886–891. [PubMed: 21964415]

Ray S, and Pyne S (2012). A Computational Framework to Emulate the Human Perspective in Flow Cytometric Data Analysis. PLoS One 7.

Rousseeuw PJ (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65.

Shapiro HM (1983). Multistation multiparameter flow cytometry: A critical review and rationale. Cytometry 3, 227–243. [PubMed: 6185284]

Sutton AJ, Abrams KR, Jones DR, and Sheldon TA Methods for Meta-analysis in Medical Research Contents Preface Acknowledgements Part A: Meta-Analysis Methodology: The Basics.

Uppal SS, Verma S, and Dhot PS (2003). Normal values of CD4 and CD8 lymphocyte subsets in healthy indian adults and the effects of sex, age, ethnicity, and smoking. Cytometry 52B, 32–36.

Weber LM, and Robinson MD (2016). Comparison of Clustering Methods for High-Dimensional Single-Cell Flow and Mass Cytometry Data. bioRxiv 47613.

Wertheimer AM, Bennett MS, Park B, Uhrlaub JL, Martinez C, Pulko V, Currier NL, Nikolich-Zugich D, Kaye J, and Nikolich-Zugich J (2014). Aging and Cytomegalovirus Infection Differentially and Jointly Affect Distinct Circulating T Cell Subsets in Humans. J. Immunol. 192, 2143–2155. [PubMed: 24501199]

Whiting CC, Siebert J, Newman AM, Du H, Alizadeh AA, Goronzy J, Weyand CM, Krishnan E, Fathman CG, and Maecker HT (2015). Large-Scale and Comprehensive Immune Profiling and Functional Analysis of Normal Human Aging. PLoS One 10, e0133627.

Wirapati P, Sotiriou C, Kunkel S, Farmer P, Pradervand S, Haibe-Kains B, Desmedt C, Ignatiadis M, Sengstag T, Schütz F, et al. (2008). Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. Breast Cancer Res. 10, R65. [PubMed: 18662380]

Yan J, Greer JM, Hull R, O ‘sullivan JD, Henderson RD, Read SJ, and Mccombe PA The effect of ageing on human lymphocyte subsets: comparison of males and females.

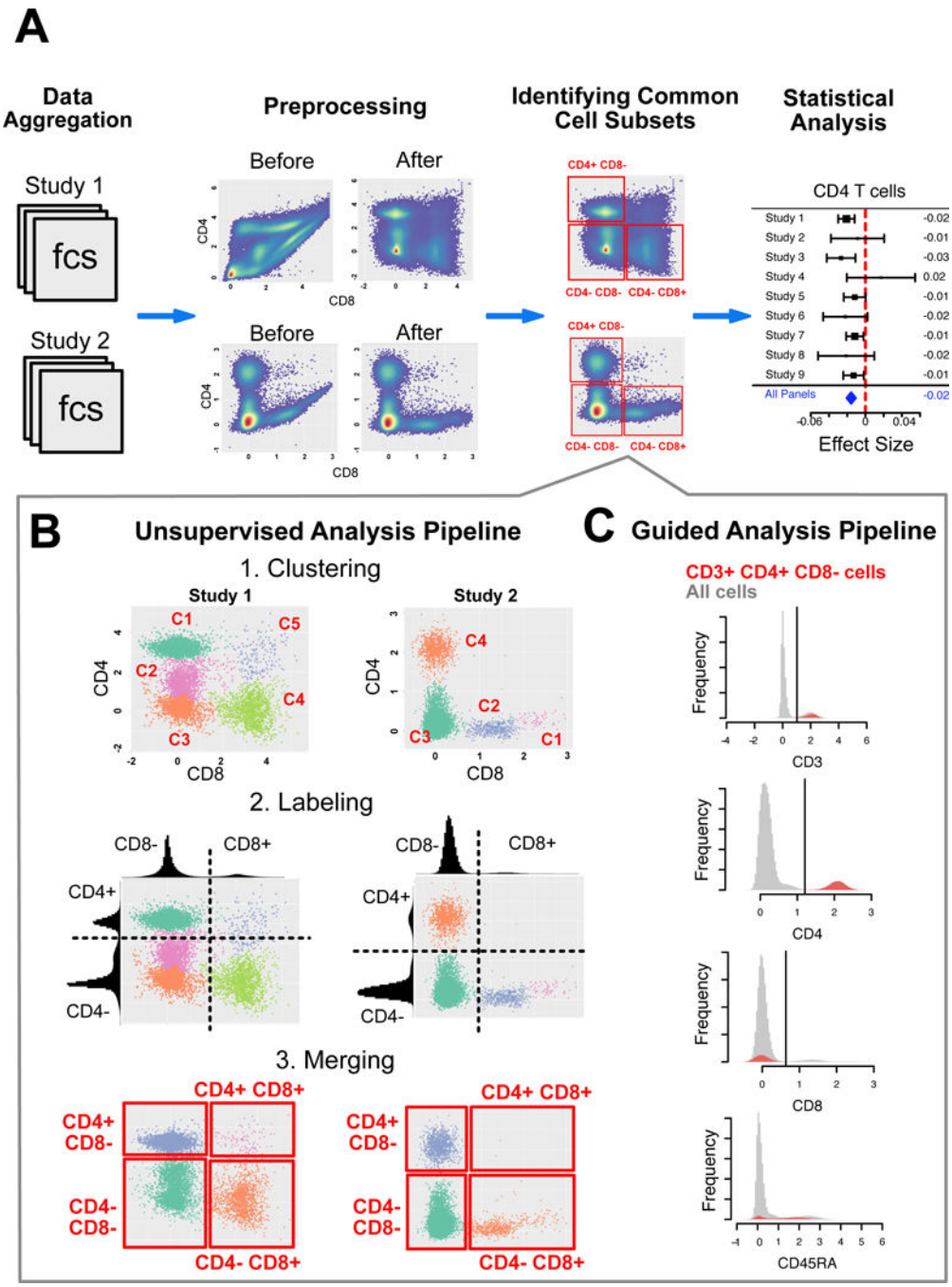

Figure 1:

MetaCyto identifies and labels common cell subsets in cytometry data across studies. (**A**) Schematic illustration of the 4 steps MetaCyto uses to perform meta-analysis of cytometry data. (**B**) Schematic illustration of the unsupervised analysis pipeline in MetaCyto. Top: Cytometry data from different studies are first clustered using a clustering method, such as FlowSOM. Middle: Each marker is bisected into positive and negative regions using the silhouette scanning method. Each identified cluster is labeled based on their position relative to this threshold. Bottom: Clusters with the same label are merged together into rectangles or

hyper-rectangles. (**C**) An example illustrating the guided analysis pipeline in MetaCyto. Each marker in the data is bisected into positive and negative regions using the silhouette scanning method. The CD3+ CD4+ CD8- cluster corresponds to cells that fall into CD3+ region, CD4+ region and CD8- region at the same time. Red histograms show the distribution of markers in CD3+ CD4+ CD8- subset. Grey histograms show the distribution of markers of all cells.

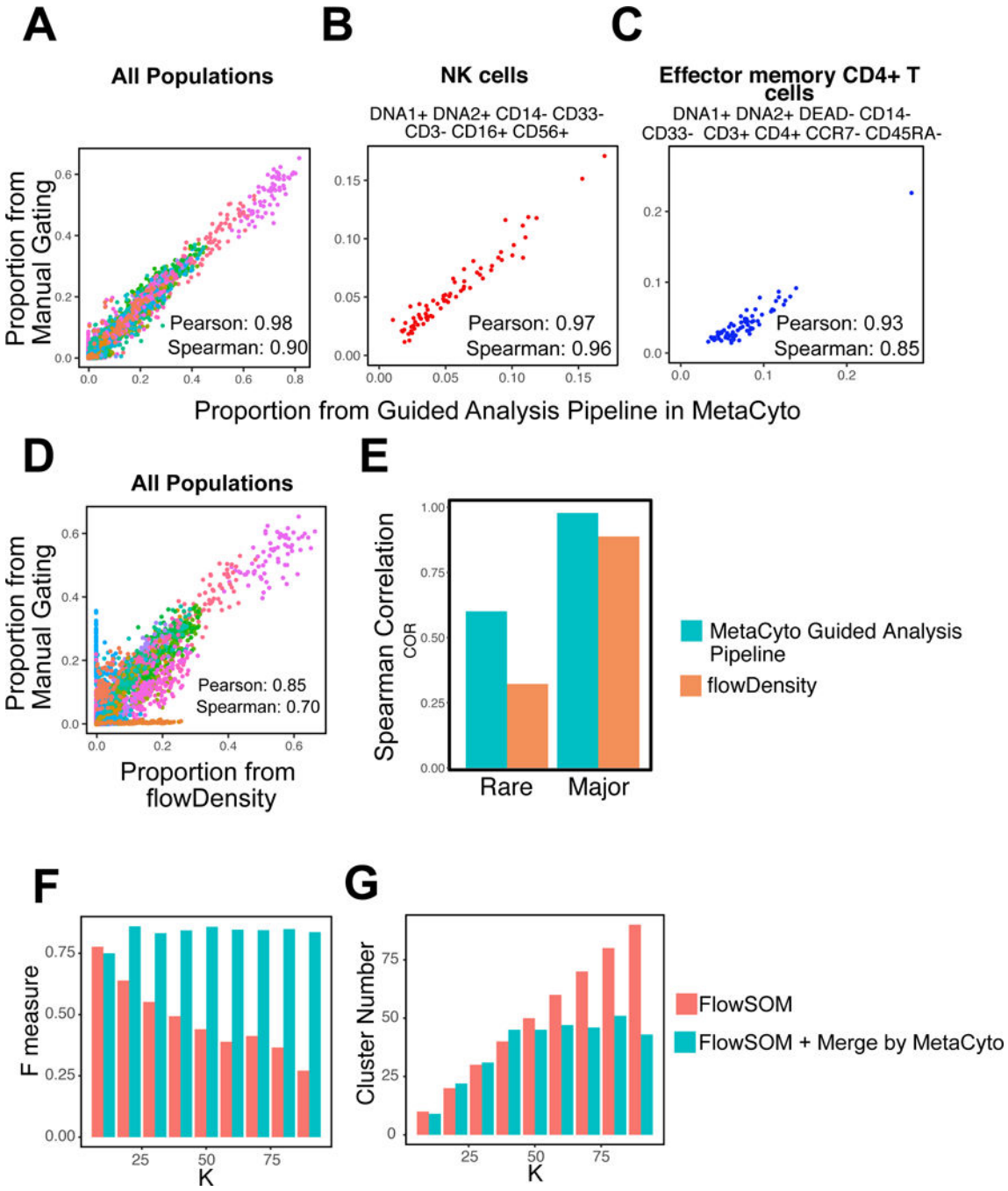

Figure 2:

Both guided and unsupervised analysis pipelines in MetaCyto accurately identify cell populations. (**A-C**) Scatter plots showing the comparison between proportions of cell types estimated by the guided analysis pipeline in MetaCyto and proportions provided by the authors of SDY478. All cell populations (**A**), Natural Killer (NK) cells (**B**), and effector memory CD4+ T cells (**C**) are included in the plots. Each dot represents the proportion of a cell type in a sample. Each color represents a cell type. (**D**) Scatter plots showing the comparison between flowDensity and manual gating. All cell populations are included. (**E**)

The 88 cell types are broken down into rare and major populations based on their mean proportion in the manual gating results. The cell types whose mean proportions are less than 2 percent are defined as rare population, the rest of the cell types are defined as major populations. Spearman correlation between MetaCyto or flowDensity's results and manual gating results are calculated to measure the performance. (**F,G**) FlowSOM is used to cluster the West Niles Virus dataset (FlowCAP WNV) with K ranging from 10 to 90 with or without the merge step in MetaCyto unsupervised analysis pipeline. F measure (**F**) and the number of clusters (**G**) are shown in the bar plots. See also Figure S2.

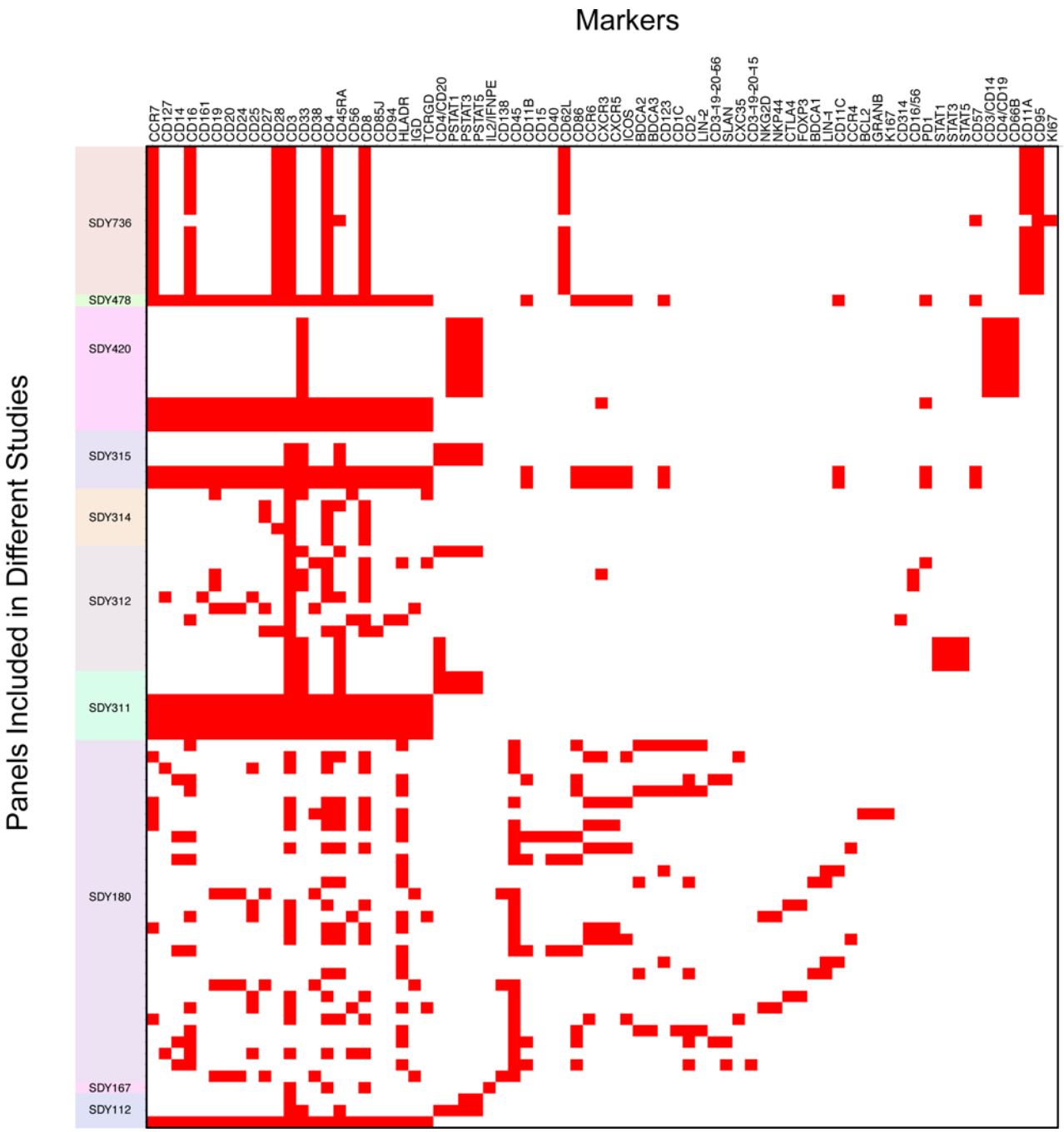

Figure 3:

Data from 10 human immunology studies includes highly heterogeneous cytometry panels. Eighty-six panels with diverse sets of markers were used in these 10 studies, with the panels represented vertically. The specific markers used are represented horizontally. Each panel is a unique antibody and fluorophores/isotope combination in a study. A red square in each grid element indicates that particular marker was used in a panel.

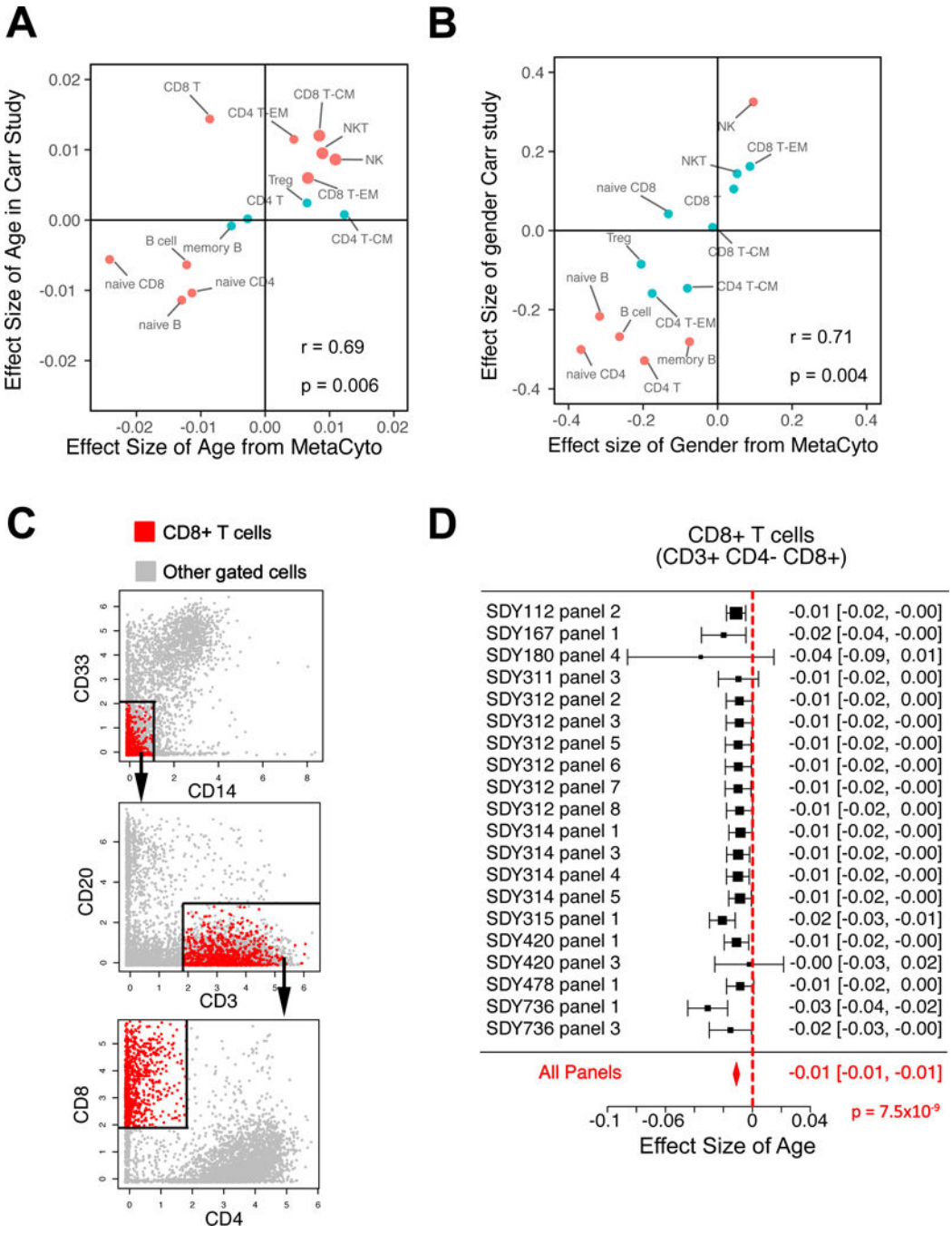

Figure 4:

Meta-analysis using MetaCyto provides consistent results between cytometry panels and confirms previous findings. (**A**-**B**) Comparison between the effect sizes of age (**A**) and gender (**B**) estimated by MetaCyto using all 86 panels, against the effect sizes estimated using the data from Carr, et al. Red dots represent significant findings by MetaCyto. (**C**) 2D plots visualizing the CD8+ T cells identified by MetaCyto. Red dots represent the cells identified by MetaCyto. Grey dots represent other cells in the parental gate. Data from SDY420 are shown as an example. Key cell lineage markers are plotted for visual

examination, not all of the markers are used for gating in MetaCyto. (**D**) A forest plot showing the effect size of ethnicity (Asian compared to White) on the proportion of CD8+ T cells in the blood. The effect sizes were estimated within each panel first, and are combined using a random effect model. In **A** and **B,** r represents the Pearson correlation; p represents the p value of r not equal to 0. In **D**, p was calculated using a random effect model.

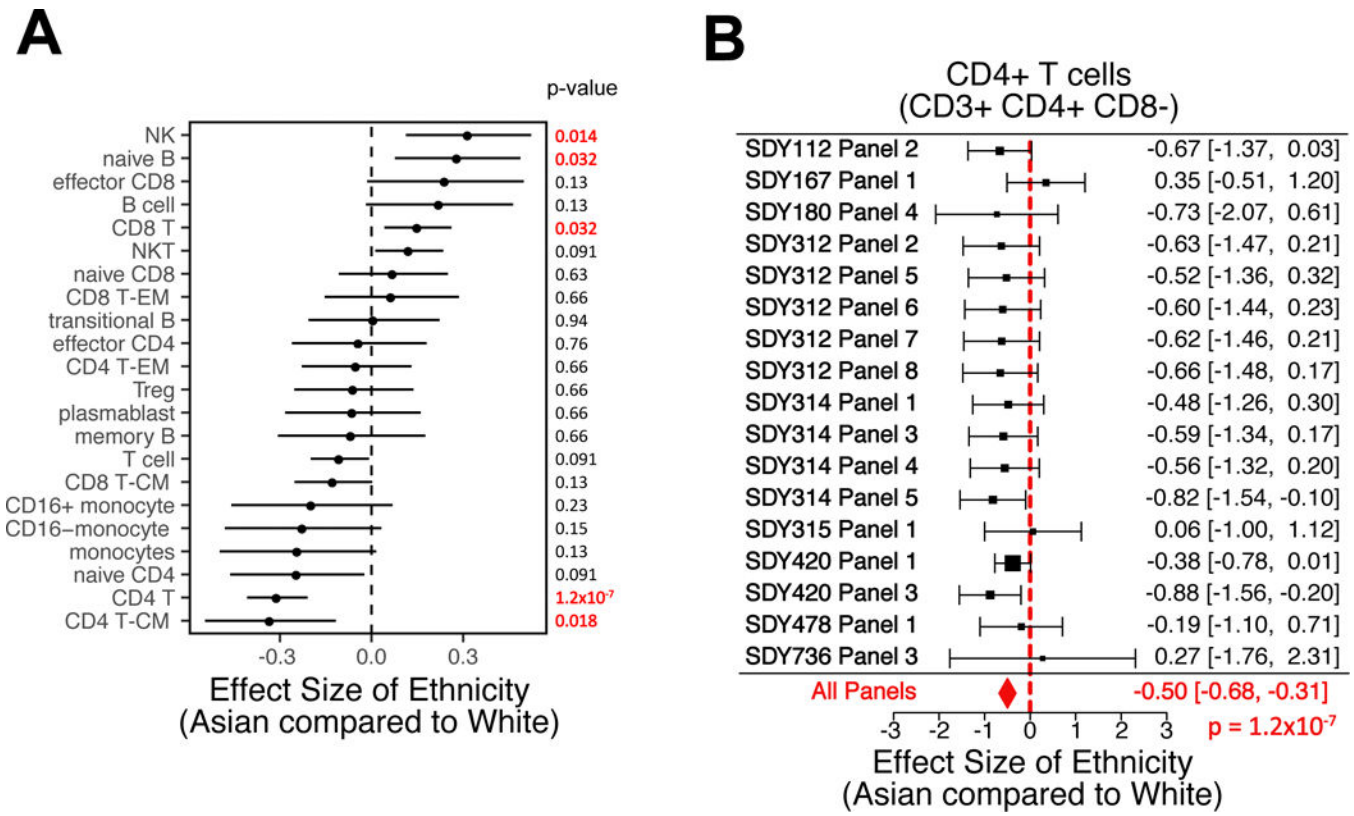

Figure 5:

Meta-analysis of cytometry data using MetaCyto identifies multiple ethnic differences in immune cells. (**A**) A plot showing the effect size of ethnicity (Asian compare to White) on the proportion of 23 cell types in blood. Dots and whiskers represent the means and 95% confidence intervals. (**B**) A forest plot showing the effect size of ethnicity (Asian compared to White) on the proportion of CD4+ T cells in the blood. The effect sizes were estimated within each panel first, and are combined using a random effect model. The p values were calculated using random effect models, adjusted using Benjamini-Hochberg correction.

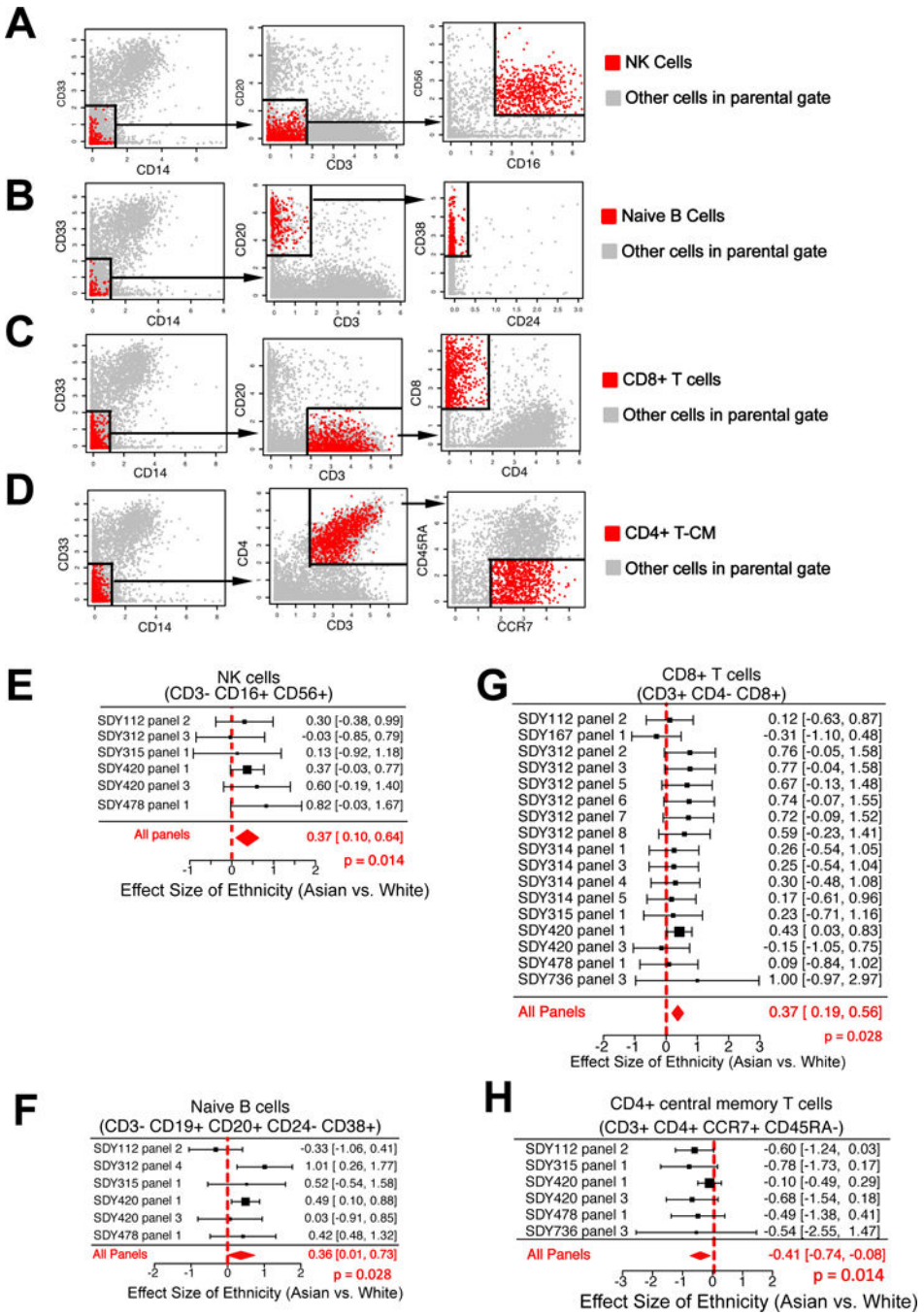

Figure 6:

Ethnic differences identified by MetaCyto are consistent across cytometry panels. (**A-D**) Representative 2D plots visualizing the cell subsets (NK Cells, Naïve B cells, CD8+ T cells and CD4+ central memory T cells) identified by MetaCyto. Red dots represent the cells identified by MetaCyto. Grey dots represent other cells in the parental gate. Data from SDY420 are shown as examples. Key cell lineage markers are plotted for visual examination, not all of the markers are used for gating in MetaCyto. (**b**) Forest plots showing the effect size of ethnicity (Asian compare to White) estimated in each cytometry

panel for Naïve B cells, CD8+ T cells and CD4+ central memory T cells. The effect sizes were estimated within each panel first, and were combined using a random effect model.